2006 CCRTS
THE STATE OF THE ART AND THE STATE OF THE PRACTICE

Content Analysis of HUMINT Reports

Dr. Matthias Hecking
FGAN/FKIE
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
Germany

Phone: +49 228 9435 576
Fax: +49 228 9435 685
hecking@fgan.de

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Content Analysis of HUMINT Reports	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Research Establishment for Applied Science (FGAN),Research Inst for Communication, Information Processing and Ergonomics (FKIE),Neuenahrer Str. 20,53343 Wachtberg-Werthhoven Germany, ,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**39**	

Content Analysis of HUMINT Reports

Dr. Matthias Hecking
FGAN/FKIE
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven, Germany
hecking@fgan.de

Abstract

The new deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of Human Intelligence (HUMINT) reports. These reports are characterized by a large topical and linguistic variety. Therefore, they are good candidates for applying techniques from computational linguistics. In this paper, the ZENON project is described, in which an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR (Kosovo Force) deployment of the Bundeswehr. The overall objective of this research is to realize a graphically navigatable Entity-Action-Network. The information about the actions and named entities are identified from each sentence and the content of the sentences are formally represented in typed feature structures. These structures can be combined and presented in a navigatable network. After a short introduction, the information extraction approach is explained. The ZENON project is described in detail. English HUMINT reports from the KFOR deployment form the basis for the development of the experimental ZENON system. These reports are used to build a specialized text micro-corpus with semantic annotations. This KFOR text corpus is described as well.

1. Introduction

On one hand, the new deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of HUMINT reports. These reports are characterized by a large topical and linguistic variety. For that reason, they are good candidates for applying techniques from computational linguistics to analyze natural languages. On the other hand, the *processing of human language* was identified as a critical capability in many future military applications (cf. [Steeneken, 1996]). Especially the *content analysis* of free-form texts is important for any information operation of the Network Centric Warfare (NCW) concept (s. [NCW, 2001], p. 5-15). We set up the research project ZENON[1], in which an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. The overall objective of this research is to realize a *graphically navigatable Entity-Action-Network*. The information about the actions and named entities are identified from each sentence and the content of the sentences are formally represented in typed feature structures. These structures can be combined and presented in the navigatable network.

The ZENON project is based on the results of the former SOKRATES[2] project. In this project we applied information extraction to the analysis of German free-form battlefield reports (cf. [Casals, 2004a], [Casals, 2004b], [Frey, 2004] , [Hecking, 2001], [Hecking, 2002], [Hecking,

[1] according to: Zenon of Citium, 336 BC - 264 BC, philosopher, founder of the Stoicism
[2] according to: Socrates, 469 BC - 399 BC, philosopher

2003a], [Hecking, 2003b], [Hecking, 2004a], [Hecking, 2004b], [Hecking, 2004c], [Schade, 2003a], [Schade, 2003b]). The SOKRATES prototype was able to process written battlefield reports (e.g., messages about hostile movements, deployments) in German. The reports were analyzed, represented in feature structures and semantically enhanced with the help of an ontology. With the SOKRATES prototype we showed the general applicability of the *Information extraction* (IE) technology for military purposes.

This paper is structured as follows. First, a short introduction into the information extraction approach is given. Then, the ZENON project is described in detail. English HUMINT reports from the KFOR deployment form the basis for the development of the experimental ZENON system. These reports are used to build a specialized text micro-corpus with semantic annotations. This KFOR text corpus is described as well.

2. Information Extraction

In the last decades various techniques for processing spoken and written natural languages were developed (e.g. speech recognizer in dictation systems, machine translation, grammar checking). IE is an engineering approach (cf. [Appelt, 1999]) for content analysis of free-form texts based on results of computational linguistics. Each IE system is tailored to a specific domain and task. IE uses a *shallow syntactic approach* (cf. [Hecking, 2003b]), i.e. that only parts of the sentences (so-called 'chunks') are processed with finite state automatons or transducers.

During the IE relevant information about the Who, What, When, etc. in natural language texts is identified, collected, and normalized (cf. [Pazienza, 1999], [Hecking, 2004a]). The relevant information is described through patterns called *templates*. These domain and task specific templates represent the meaning of the relevant information. During the IE task the templates are filled with the extracted information. One possibility to realize the templates is to use *typed feature structures* (cf. [Hecking, 2004b]). Therefore, IE can be seen as the process of normalizing free-form text into a defined semantic structure.

To realize an IE system, language-specific resources (lexicon, grammar) and appropriated parsing software are necessary.

In order to achieve robust and efficient IE systems, domain knowledge must be integrated and shallow algorithms must be used. The domain knowledge is tightly integrated with the language knowledge, e.g., the name 'Leopard' in the lexicon has the categorical information 'tank'. This association between words and semantic information is domain-specific and has to be change for other applications.

The IE process itself is divided into sub steps. After tokenizing the text, the sentence boundaries must be identified. Then, the morphological component identifies the word stems, the abbreviation, and detects the syntactic information (e.g., grammar case and gender). After this, the chunk parsing with transducers selects parts of the natural language text that are relevant for

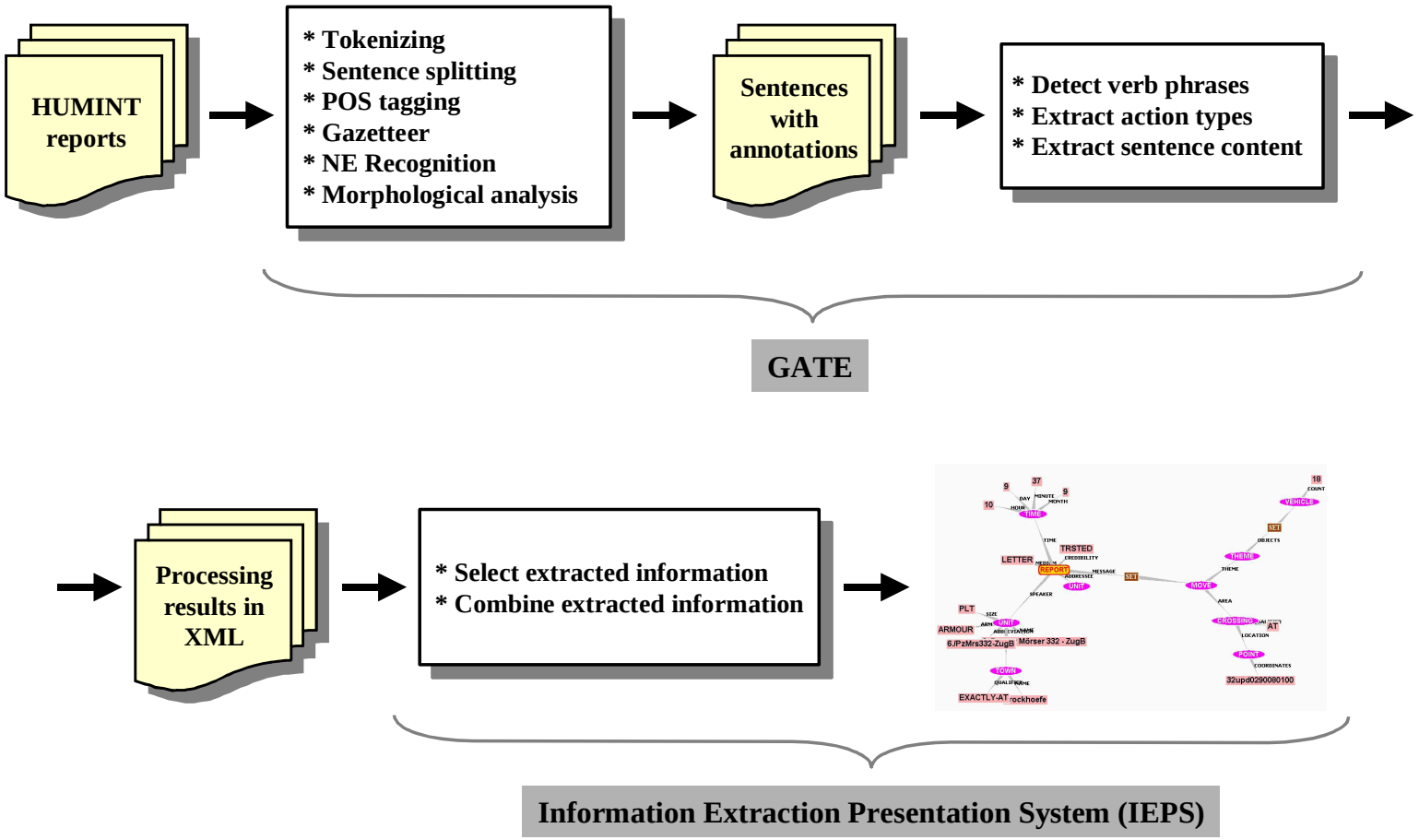


Figure 1: The ZENON processing chain

the specific information extraction task. The chunks are then used to instantiate the templates, which represent the action/event descriptions. They are the result of the IE process.

The IE is used as the core natural language processing technique in the ZENON project.

3. The ZENON Project

Outline

Starting with English HUMINT reports (and a list of the city names) from the KFOR deployment of the German Federal Armed Forces we have realized in our ZENON project a prototype that is able to do a (partial) content analysis of these reports (cf. [Hecking, 2005a]). The content of these KFOR reports are from a wide spectrum. Apart from descriptions of conflicts between ethnic groups, tensions between political parties, information about infrastructure problems, etc. there are also reports, which concern individuals or other entities. Statements of the form ´A meets B´, ´A marries C´, ´A shoots B´, etc. contains information about activities/events and involved entities. This information, completed with location and time data, is combined into a graphically navigatable Entity-Action-Network (e.g.; with a person in the center of the network). The intelligence analysts can use this network to navigate through the content of the reports.

Since most of the reports are in English, GATE (General Architecture for Text Engineering, cf. [Cunningham, 2002]) was selected as the used toolbox. GATE is an architecture, a free open source framework (SDK) and graphical development environment for Natural Language Engineering and offers a lot of tools, which are used to realize the natural language processing parts of the ZENON prototype (e.g., morphological analyzer, part-of-speech (POS) tagger, pre-defined transducer to recognize English verbal phrases, chunk-parsing). The functionality to select and combine the extracted information from different sentences and different reports is realized by the *Information Extraction Presentation System* (IEPS). IEPS is a graphical tool to visualize information extracted from free-form texts.

In Figure 1 the ZENON processing chain is shown. HUMINT reports are fed into the first sub-component. In this component the natural language text is tokenized (i.e., find words, numbers, etc.), the sentence boundaries are detected, the part-of-speech (i.e., whether it's a noun, a verb, etc.) is determined, simple names of cities, regions, military organizations etc. are annotated (through the Gazetteer), named entities (i.e., complex names of e.g. political organizations, person names, etc.) are recognized and a morphological analysis is done. The result of this sub-component are the annotated sentences of the reports. The second sub-component uses these annotations to extract the action type (e.g., 'kill') starting with the verb of the sentence. If the action type is determined the other parts of the sentence (e.g., subject, object, time expressions) are located and formally represented in *typed feature structures*. These structures are coded in XML (Extensible Markup Language) format and represent the output of the natural language part of the ZENON prototype. In the third sub-component (IEPS) the extracted content of different reports can be combined and selected according to predefined XSLT (Extensible Stylesheet Language Transformation) sheets. The result of the analysis can be navigated interactively.

Extraction of Named Entities

An important processing step during the natural language processing is the recognition of the domain- and application-specific named entities. In the ZENON prototype transducers for the

recognition of the following named entities were developed: *City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time* and *Title*. An example is shown in Figure 2.

```
Rule: PersonName1
////////////////////////////////////////////////////
// Recognizes: "Mr. Bedredin SHEHU", "Mrs SHEHU"
// Output:     TempPerson{title, firstName, lastName,
                        gender, 'PersonName1'}

 (
     // Titel
     (
          ({PersonTitle}):title
          ({Token.string == "."})?
     )

     // First name
     (
          ({Lookup.majorType == person_first}):firstName
     )?

     //Last name
     ({Token.category == NNP,
       Token.orth == allCaps}):lastName
     (
          {Token.string == "-"}
          {Token.category == NNP}
     )?
):person

-->

{...}
```

Figure 2: Named entity recognizer 'PersonName1'

Extraction of Verb Phrases, Action Types and Sentence Content

GATE offers various transducers to recognize the English verb groups. We have adapted and extended these transducers to fit our application. In addition to finite and non-finite verbal phrases also modal verb phrases, participles and special composed verb expressions are recognized. Figure 3 shows an example.

```
Rule: FVGPrePerPasNeg
////////////////////////////////////////////////////
//Recognizes:  Present Perfect Passive Negative:
               "hasn't been eaten"
//Pattern:     (has | have) not been VBN
//Output:      VG{adverb, infinitive, neg='yes',
               tense='PrePer', type='FVG', voice='passive'}

(
     (
          {Token.string == "has"} |
          {Token.string == "have"}
     )
     (NEGATION)
     {Token.string == "been"}
     (ADVS):adverb
     ({Token.category == VBN}):verb
):x

-->

{...}
```

Figure 3: Verb phrase transducer 'FVGPrePerPasNeg'

Based on the recognized verb groups different action types can be detected (e.g., from the infinitive of 'murder', 'kill', 'decapitate', … the action class 'kill'). After detecting the action type the verb phrase and other parts of the sentence must be combined. In the ZENON project we use the *Semantic Frames* from the FrameNet project (cf. [FrameNet]) to realize this combination. Semantic frames are schematic representations of situation types (eating, killing, spying, classifying, etc.) together with lists of the kinds of participants, objects, and other conceptual roles that are seen as components of such situations. These semantic arguments are called the frame elements of the frame. Figure 4 shows an example. The core (must exist) frame elements for the frame 'killing' are CAUSE or KILLER and VICTIM. In the example sentence 'John' fills the role KILLER and 'Martha' fills the role VICTIM.

Semantic Frame 'killing':	A KILLER or CAUSE causes the death of the VICTIM.
Core frame elements:	CAUSE, KILLER, VICTIM
Non-core frame elements:	DEGREE, DEPICTIVE, INSTRUMENT, MANNER, MEANS, PLACE, PURPOSE, REASON, RESULT, TIME
Example sentence:	[John KILLER] DROWNED [Martha VICTIM].

Figure 4: Frame 'killing'

Associated with each Semantic frame are examples with typical syntactic realization of the frame elements. These examples and examples from the KFOR reports form the basis to construct the transducers, which produce the sentence content.

During the processing, the associated Semantic Frame is inferred from the detected action type. With the identified Semantic Frame the core and non-core frame elements are give. Recognized named entities, POS tagging and expressions from the sentences are used to fill in the frame elements.

For example, after processing

John Mueller and four other persons were killed in an explosion incident in GOSTIVAR area.[3]

the named entities for the person, the city and the number (see Figure 5) and the verb group (see Figure 6) were produced. This information together with miscellaneous language material from the sentence is used to produce the content representation of the whole sentence (see Figure 7). The type of the sentence is 'kill'. For the victims there are two pieces of information. The value of ':victim' is the identified person. The name of the victim doesn't span the whole expression in front of the verb phrase. Therefore, there is more information about possible victims in this sentence. For not loosing any information, the whole expression is stored in ':victimAll'. The cause of the killing-event is stored in ':causeAll'. The information where the killing took place is given through ':place' and ':placeAll'.

```
{                       {                       {
   :type=:person,          :type=:city,            :type=:number,
   :firstName=John,        :name= GOSTIVAR,        :value=4,
   :lastName=Mueller,      …                       …
   …                    }                       }
}
```

Figure 5: Recognized named entities (abbreviated)

```
{
   :type=:vg,
   :infinitive=kill,
   :typeVG=FVG,
   :tense=SimPas,
   :voice=passive,
   :rule=FVGSimPasPas
}
```

Figure 6: Recognized verb group (abbreviated**)**

[3] The names are not the real ones.

```
{
    :type=kill,

    :verb=
        {
            :type=:vg,
            :infinitive=kill,
            :typeVG=FVG,
            :tense=SimPas,
            :voice=passive,
            :rule=FVGSimPasPas
        },

    :victim=
        {
            :type=:person,
            :firstName=John,
            :lastName=Mueller,
            :start=534,
            :end=546,
            :rule=PersonName2
        },

    :victimAll= John Mueller and four other persons,

    :causeAll=an explosion incident,

    :place=
        {
            :type=:city,
            :name=GOSTIVAR,
            :rule=City
        },

    :placeAll= GOSTIVAR area,

    :sentenceContent=John Mueller and four other persons were killed in an
                        explosion incident in GOSTIVAR area.,

    :start=534,
    :end=624,
    :rule=killAction2

}
```

Figure 7: Formal representation of sentence content

Information Extraction Presentation System

The natural language processing module of the ZENON prototype creates for each sentence in each KFOR report a formal representation of the content. This contains information pieces about activities, events, entities, times and places. These pieces are now put together according to specific analysis requirements (e.g., all information about a specific person). The result of this recombination is a *graphically navigatable Entity-Action-Network*. The intelligence analyst can use this network for faster access the important information from the used set of reports.

In the ZENON prototype the above describe functionality is realized by the *Information Extraction and Processing System* (IEPS, cf. [Casals, 2005]). IEPS is a graphical software tool (see Figure 8) for visualizing information typically extracted from free-form texts by a natural language processing system. Additionally, it offers a framework to organize all the files being employed during the processing in user-defined scenarios and to activate the IE process. IEPS represents extracted information by means of interactive graphs. The results of the IE system can be filtered in different ways, either by hiding unimportant information, or by combining results from different reports and sentences. This is done by using XSLT. The visual interface is based on TouchGraph (cf. [TouchGraph]).

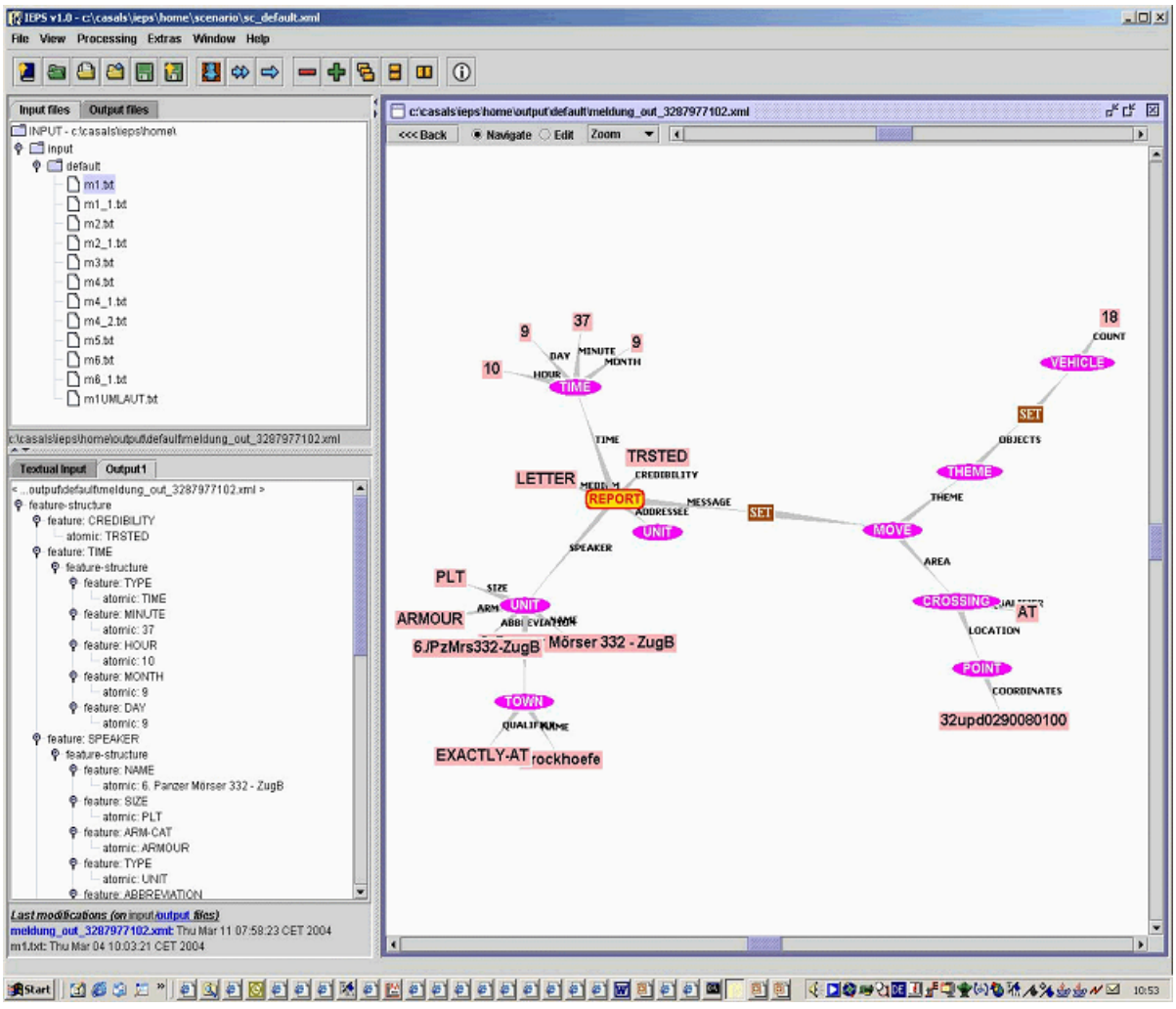


Figure 8: Information Extraction and Processing System (IEPS)

4. The KFOR Corpus

4,498 military reports (mostly in English) from the KFOR deployment of the German Federal Armed Forces were used for the realization of the ZENON prototype. From these reports 800 were manually annotated and form the *KFOR Corpus*[4]. This corpus is a specialized micro text corpus (cf. [McEnery, 2001, p. 191]). The corpus covers 886,000 tokens and contains the annotations in different layers (cf. [Hecking, 2005b]). The following layers are available:

- Original markups: In this layer those parts of the message are annotated, which are already formatted (e.g. addressee, topic, source) .
- Token: This layer contains the annotations, which are supplied by the tokenizer and the POS tagger.
- Gazetteer: In this layer those expressions are annotated, which were identified over lists of names (e.g., first names, city names).
- Sentence: These annotations refer to sentences and begin and end markers of comments.
- Named entities: City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time and Title.
- Verb Group: The verbal phrases are annotated.

During the creation of the corpus a first version of the annotations were produced automatically. These annotations were then checked manually and corrected. For both working-steps GATE was used. The corpus is represented in

- the GATE-specific format,
- the GATE-specific format in XML, and
- the ANC (American National Corpus) stand-off annotation format.

For the following purposes the KFOR corpus is used:

1. It represents the basis for the construction of the IE component. The lexicon and the transducers are optimized towards the corpus.
2. The performance of the ZENON IE can be quantitatively evaluated relative to the KFOR corpus.
3. The KFOR corpus can be used for other research objectives (e.g., complexity of nominal phrases, word sense disambiguation, machine learning of grammatical structures, etc.).

[4] Since the KFOR corpus is classified, it is not freely available.

4. Conclusion

In this paper, the ZENON project was presented. In this project an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. First, a short introduction into the information extraction approach was given. Then, the ZENON project was described in detail. The KFOR text corpus was mentioned as well.

A first version of the prototype was constructed. It is able to process sentences of action type 'kill'. This will be extended to other action types. On the natural language processing side using WordNet will extend the processing capabilities.

References

[Appelt, 1999] Appelt, D. & Israel, D. *Introduction to Information Extraction Technology*. Stockholm: IJCAI-99 Tutorial, 1999, http://www.ai.sri.com/~appelt/ie-tutorial/.

[Casals, 2004a] Casals Elvira, X. *Project SOKRATES: Interaction with the LC2IEDM Database*. FKIE-Bericht, Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., to appear, 2004.

[Casals, 2004b] Casals Elvira, X. *Project SOKRATES: Processing of Headers for the Information Extraction Component*. FKIE-Bericht, Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., to appear, 2004.

[Casals, 2005] Casals Elvira X., Hecking, M.. IEPS: *A Framework to Manage and to Visualize Information Extraction Results*. Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN), Technischer Bericht FKIE/ITF/2005/2, September 2005.

[Cunningham, 2002] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[FrameNet] http://framenet.icsi.berkeley.edu/index.php (11.1.2006).

[Frey, 2004] Frey, M. L. & Schade, U. *Modular Framework for Military Report Processing*. FKIE-Bericht. Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., FKIE-Bericht Nr. 73, 2004.

[Hecking, 2001] Hecking, M. *Natural Language Access for C2 Systems*. Paper presented at the RTO IST Symposium on „Information Management Challenges in Achieving Coalition Interoperability“, held in Québec, Canada, 28-30 May 2001, and published in RTO MP-064.

[Hecking, 2002] Hecking, M. *Analysis of Spoken Input to C2 Systems*. In: Proceedings of the 7th International Command and Control Research and Technology Symposium (ICCRTS), Québec City, Canada, 2002.

[Hecking, 2003a] Hecking, M. *Information Extraction from Battlefield Reports*. In: Proceedings of the 8th International Command and Control Research and Technology Symposium (ICCRTS), Washington, DC, U.S.A., 2003.

[Hecking, 2003b] Hecking, M. *Analysis of Free-form Battlefield Reports with Shallow Parsing Techniques*. Paper presented at the RTO IST Symposium on „Military Data and Information Fusion“, held in Prague, Czech Republic, October 20-22, 2003.

[Hecking, 2004a] Hecking, M. *Informationsextraktion aus militärischen Freitextmeldungen*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 74, 2004.

[Hecking, 2004b] M. Hecking. *How to Represent the Content of Free-form Battlefield Reports*. In: Proc. of the 2004 Command and Control Research and Technology Symposium (CCRTS) "The Power of Information Age Concepts and Technologies", June 15-17, 2004, San Diego, California.

[Hecking, 2004c] M. Hecking. *Improve Interoperability by Formalizing the Natural Language Parts of Military Messages*. In: Proc. of the Information Systems Technology Panel Symposium "Coalition C4ISR Architectures and Information Exchange Capabilities", September 27-29, 2004, The Hague, The Netherlands.

[Hecking, 2005a] M. Hecking. *Domänenspezifische Informationsextraktion am Beispiel militärischer Meldungen*. In: A.B. Cremers, R. Manthey, P. Martini, V. Steinhage (Hrsg.) "INFORMATIK 2005", Band 2, Lecture Notes in Informatics, Volume P-68, Bonn, 2005.

[Hecking, 2005b] M. Hecking. KFOR-Korpus – Annotierungsvorschrift. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), unpublished, 2005.

[McEnery, 2001] T. McEnery and A. Wilson. *Corpus Linguistics*. Edinburgh University Press, Edinburgh, 2nd edition, 2001.

[NCW, 2001] Department of Defense. *Network Centric Warfare – Report to Congress*. 27 July 2001.

[Pazienza, 1999] Pazienza, M. T. (ed.) *Information Extraction*. Berlin, 1999.

[Schade, 2003a] Schade, U. *Ontologieentwicklung für Heeresanwendungen*. Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN), FKIE-Bericht Nr. 57, 2003.

[Schade, 2003b] Schade, U. *Towards an Ontology for Army Battle C2 Systems*. In: Proceeding of the 8th ICCRTS, 2003.

[Steeneken, 1996] Steeneken, H. J. M. *Potentials of Speech and Language Technology Systems for Military Use: an Application and Technology Oriented Survey*. NATO, Technical Report, AC/243(Panel 3)TP/21, 1996.

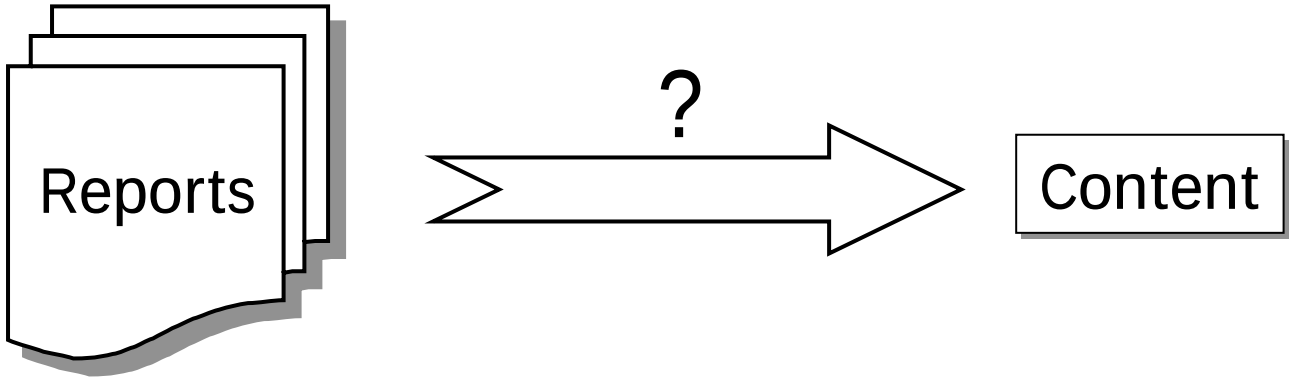
[TouchGraph] http://www.touchgraph.com (11.1.2006).

Content Analysis of HUMINT Reports

Dr. Matthias Hecking
Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN)
Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE)
Abtl. Informationstechnik und Führungssysteme (ITF)
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
hecking@fgan.de

Content

- 1. Introduction
- 2. Information Extraction
- 3. Project ZENON
- 4. KFOR Corpus



- General problem: There are a lot of natural language texts (military reports, emails, web pages, scientific reports, documents, ...) which can't be evaluated due to missing specialists.
- Which technical possibilities exist of automating the content extraction?

➔ Practical approach: Information Extraction (IE)

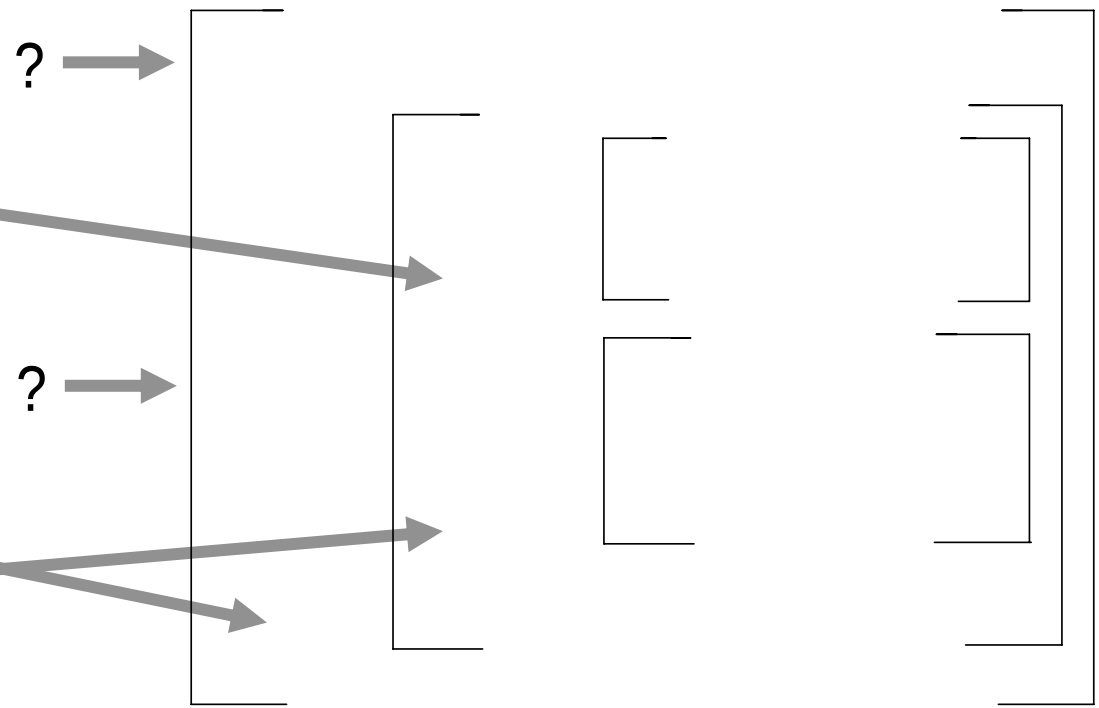
- Specific problem: content extraction of HUMINT reports
- ZENON project: The overall objective is to realize an experimental system for (partial) *content extraction of HUMINT reports* from the KFOR deployment of the Bundeswehr and to realize a possibility to evaluate the formal representation of the content.
- For the realization of the IE module approx. 4000 English HUMINT reports are available.

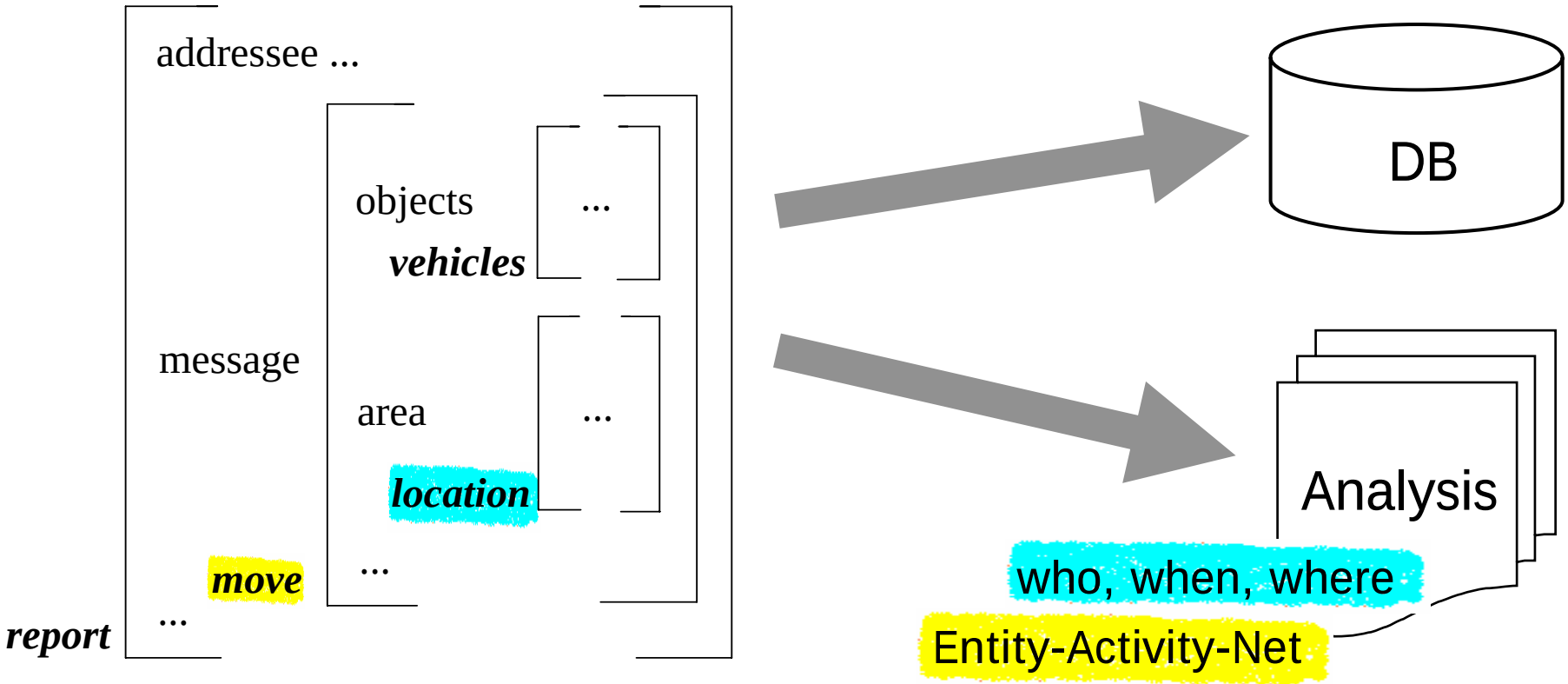
- For the realization the toolbox GATE is used.
- The ZENON prototype was integrated into the "CliCC" system.
- "CliCC" is a German contribution for CWID 2006.
- For the evaluation of the ZENON IE the KFOR text corpus was developed.

- Information extraction (IE) is the task of identifying, collecting and normalizing information from natural language text.
- Relevant information about the Who, What, When, etc. is looked for.
- The information of interest is described through domain-specific lexicon rules and patterns called *templates*.
- During the IE task these templates are filled with the collected information.
- The templates are domain and task specific, i.e. for each new task and domain they must be newly created.

„… in TUZLA:
10.20 a.m.
18 vehicles
- 8 with attached ZIS–3 and 1 with attached T 12 -
march

at road crossing (CQ 072368) south of MILESKIJ (CQ 0737) to the north.“

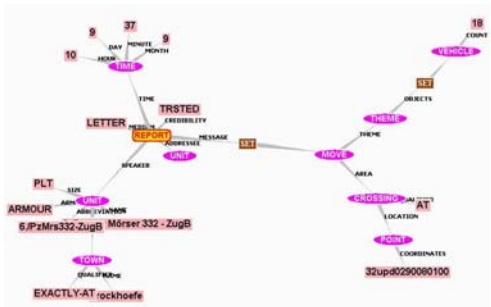


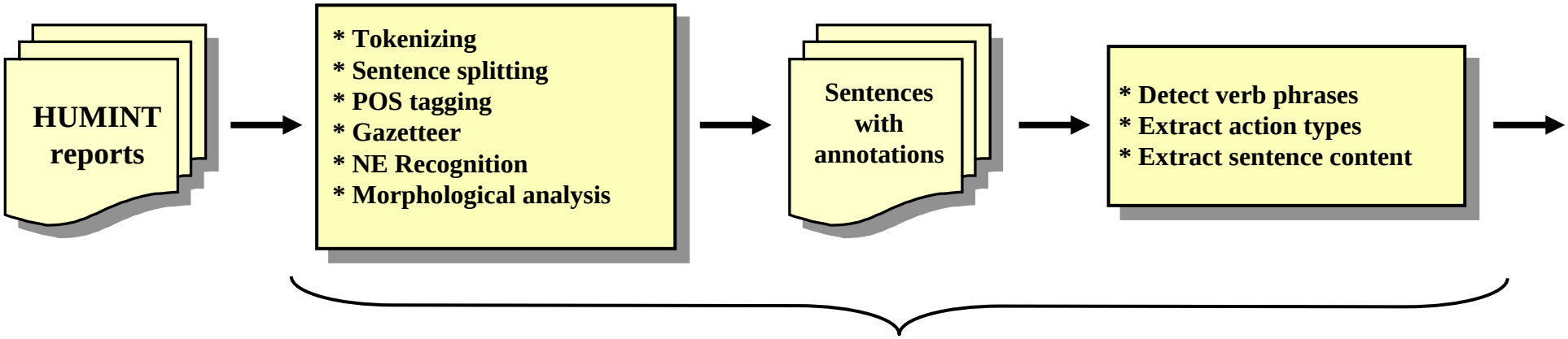

addressee ...
objects
...
vehicles
message
area
...
location
...
move
...
report
DB
Analysis
who, when, where
Entity-Activity-Net

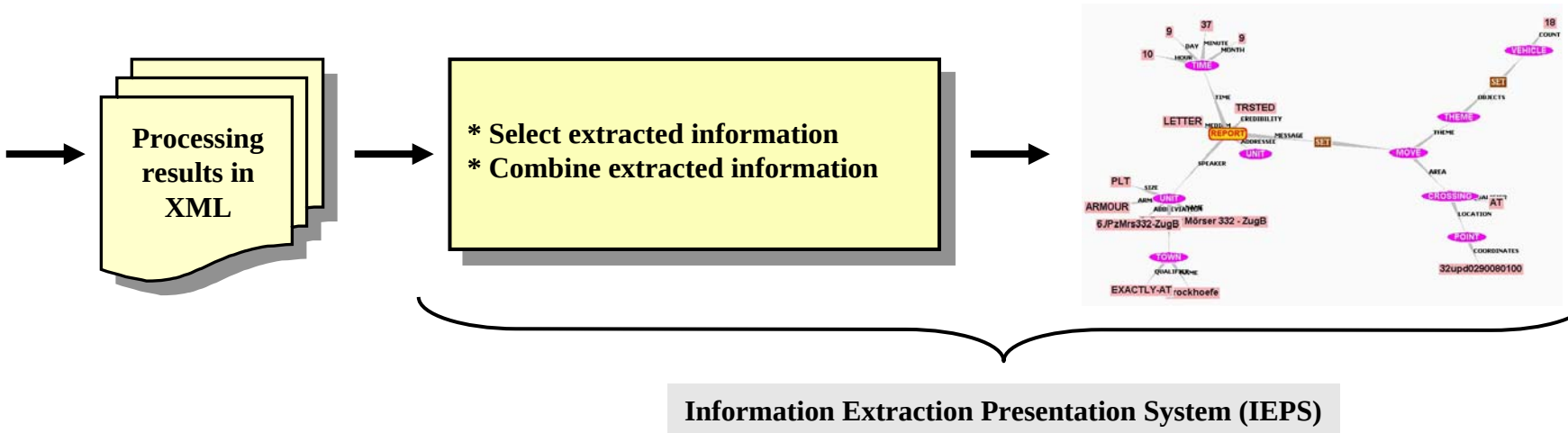
- Technical basis of the information extraction:
  - extensive linguistic knowledge
    - general and application-specific lexica
    - general and/or application-specific phrase grammars
    - general (and/or application-specific) clauses and sentence grammars
  - cascaded transducers, i.e. finite state automaton that reads from the input and writes to the output
  - only application-relevant parts of the texts are analyzed through transducers (shallow parsing techniques)

The experimental ZENON project:

- The Information Extraction (IE) technology is used for the content extraction.
- The information about the actions and named entities are identified from each sentence and the content of the sentences are formally represented in typed feature structures.
- These structures can be combined and presented in a graphically navigatable *Entity-Action-Network*.




HUMINT reports
* Tokenizing
* Sentence splitting
* POS tagging
* Gazetteer
* NE Recognition
* Morphological analysis
Sentences with annotations
* Detect verb phrases
* Extract action types
* Extract sentence content
GATE


Processing results in XML
* Select extracted information
* Combine extracted information
TIME
LETTER
TRSTED
REPORT
UNIT
MOVE
VEHICLE
THEME
CROSSING
POINT
TOWN
ARMOUR
PLT
6/PzMrs332-ZugB
Mörser 332 - ZugB
EXACTLY-AT
rockhoefe
32upd0290080100
Information Extraction Presentation System (IEPS)

GATE:

- "is one of the most widely used human language processing systems in the world."
- "comprises an architecture, framework (or SDK) and graphical development environment ..."
- "... has been under construction in Sheffield since 1995."
- "The system has been used for many language processing projects; in particular for Information Extraction in many languages."
- "GATE is funded by the Engineering and Physical Sciences Research Council (EPSRC), the EU and commercial users."
- http://gate.ac.uk/

- Chunk parsing of named entities (NE): *City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time* and *Title*
- Example:

```
Rule: PersonName1
 (
     (      ({PersonTitle}):title ({Token.string == "."})?          )
     (      ({Lookup.majorType == person_first}):firstName       )?
     ({Token.category == NNP, Token.orth == allCaps}):lastName
     ({Token.string == "-"}{Token.category == NNP})?
):person
-->
{...}
```

Determine action types:

- extraction of verb phrases (modal verb phrases, participles, special composed verb expressions)
- mapping from recognized verb groups to action types (e.g., from the infinitive of 'murder', 'kill', 'decapitate', ... to action type 'kill').

Sentence analysis:

- The basic structures for the semantic sentence analysis are given by the FrameNet-Project.
- Semantic roles are specified for the lexical units (frames, verbs).
- Example: Frame *Killing*
  - Def.: A KILLER or CAUSE causes the death of the VICTIM.
  - Roles: CAUSE, KILLER, VICTIM, DEGREE, INSTRUMENT, ...
  - Sentence: [John Mueller and four other persons **VICTIM**] were killed in [an explosion incident **CAUSE**] in [GOSTIVAR area **PLACE**].
  - Formal representation ...

Dr. M. Hecking

```
{
    :type=kill,
    :verb={:type=:vg, :infinitive=kill, :typeVG=FVG, :tense=SimPas, ...},
    :victim={:type=:person, :firstName=John, :lastName=Mueller, ...},
    :victimAll= John Mueller and four other persons,
    :causeAll=an explosion incident,
    :place={:type=:city, :name=GOSTIVAR, …},
    :placeAll= GOSTIVAR area,
    :sentenceContent=John Mueller and four other persons were killed …,
    :start=534,
    :end=624,
    :rule=killAction2
}
```

Information Extraction Presentation System (IEPS):

- Graphical presentation of the extracted information (typed feature structures)
- Scenario: Name, sets of input and output files (XML coded feature structures), filter.
- Filter: Description of the transformation with XSLT

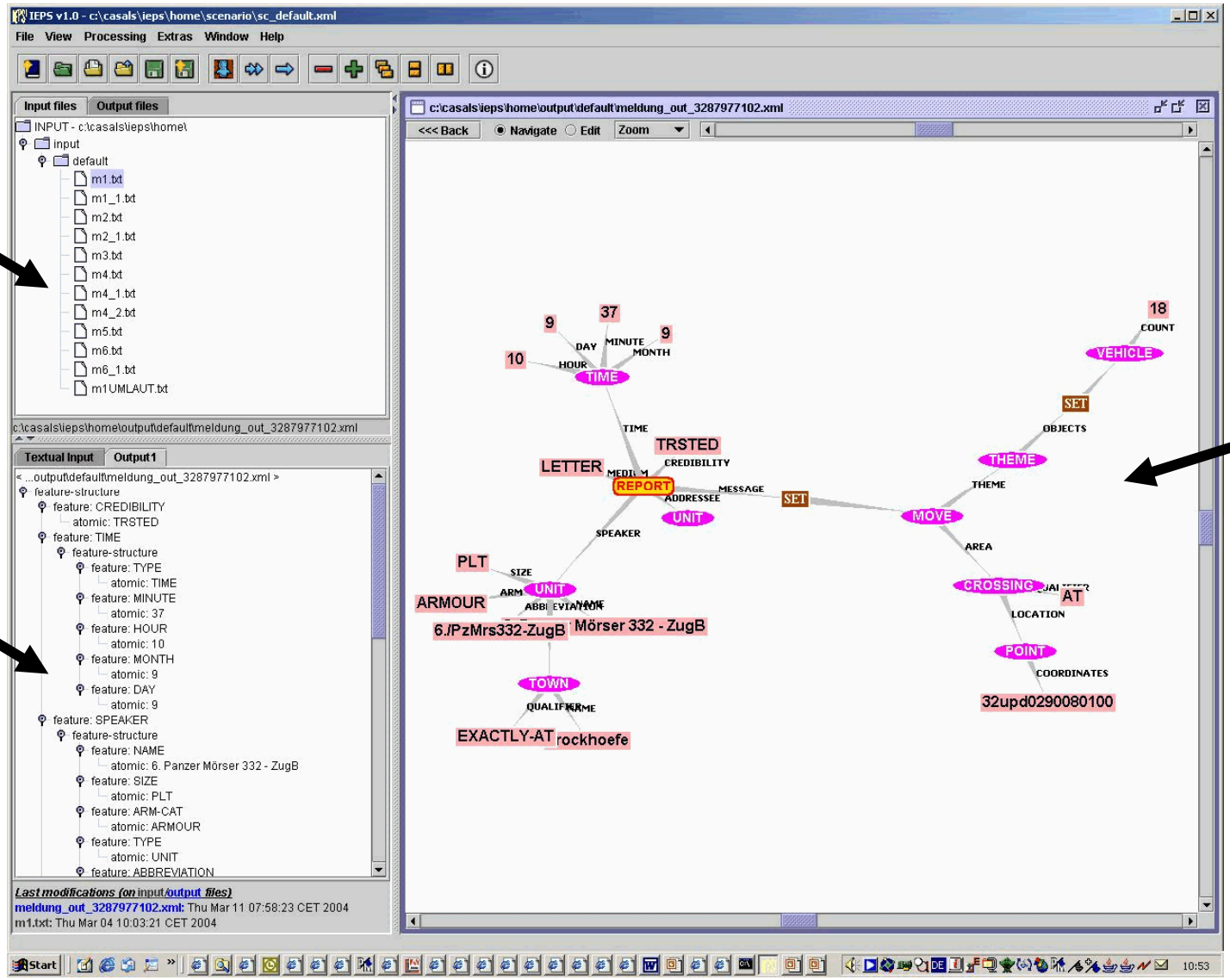
3 | ZENON Project: IEPS - II

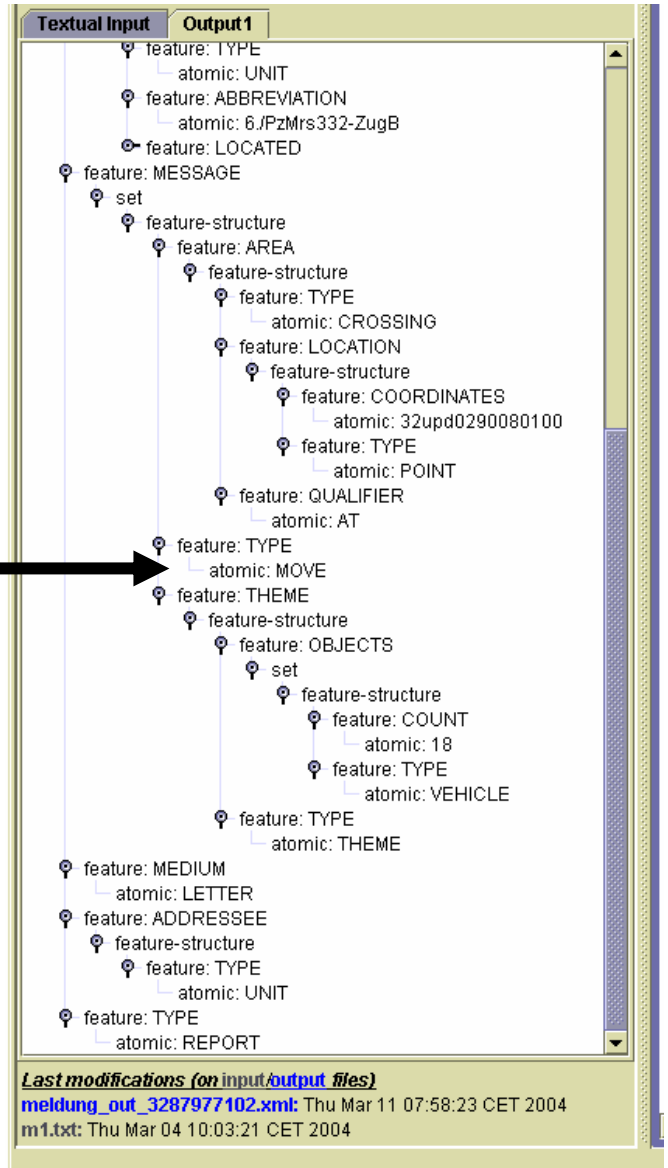
Dr. M. Hecking

input files

formal output

graphical output



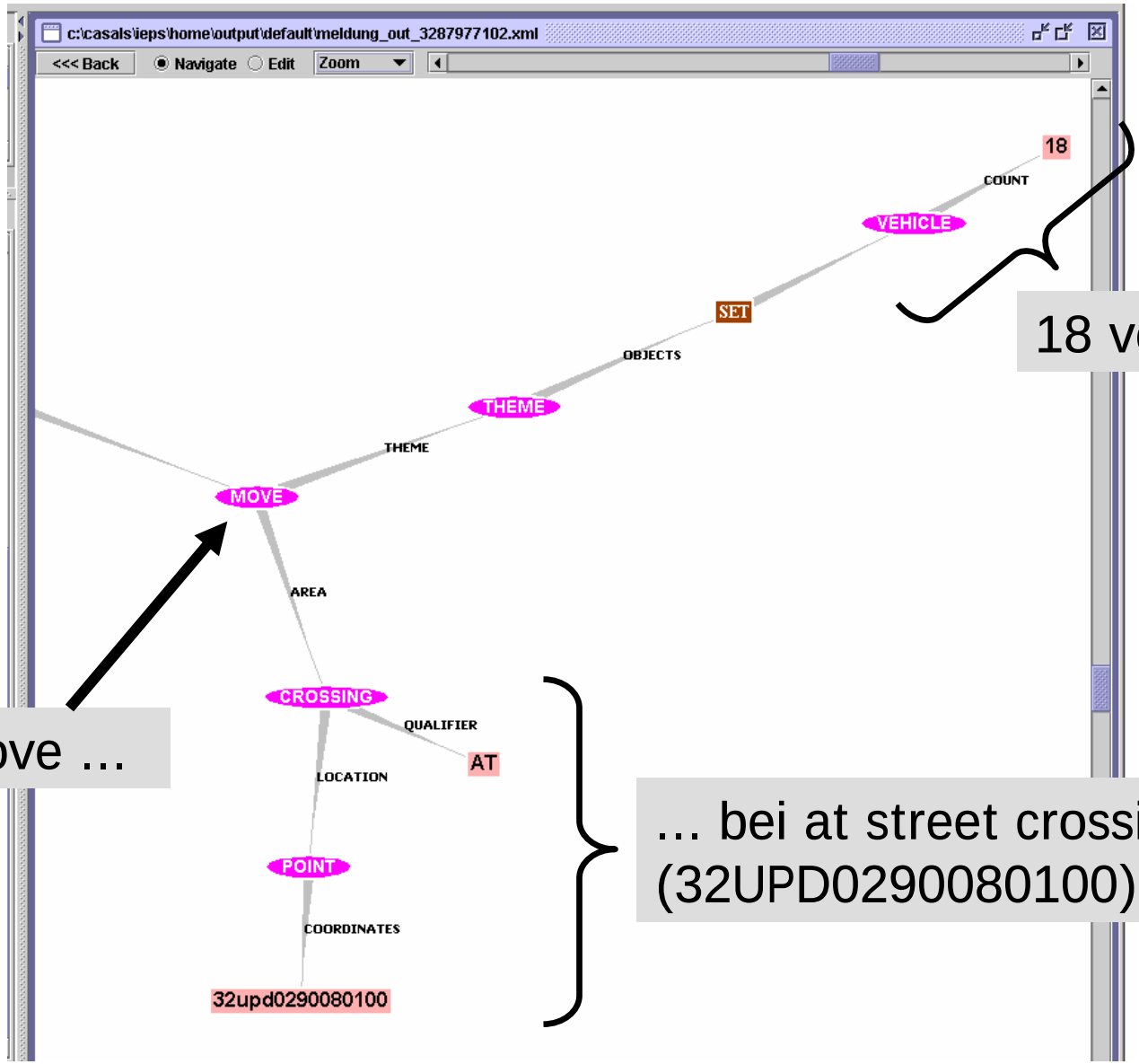


... move ...

... at street crossing (32UPD0290080100)

18 vehicles

Dr. M. Hecking



18 vehicles

... move ...

... bei at street crossing (32UPD0290080100)

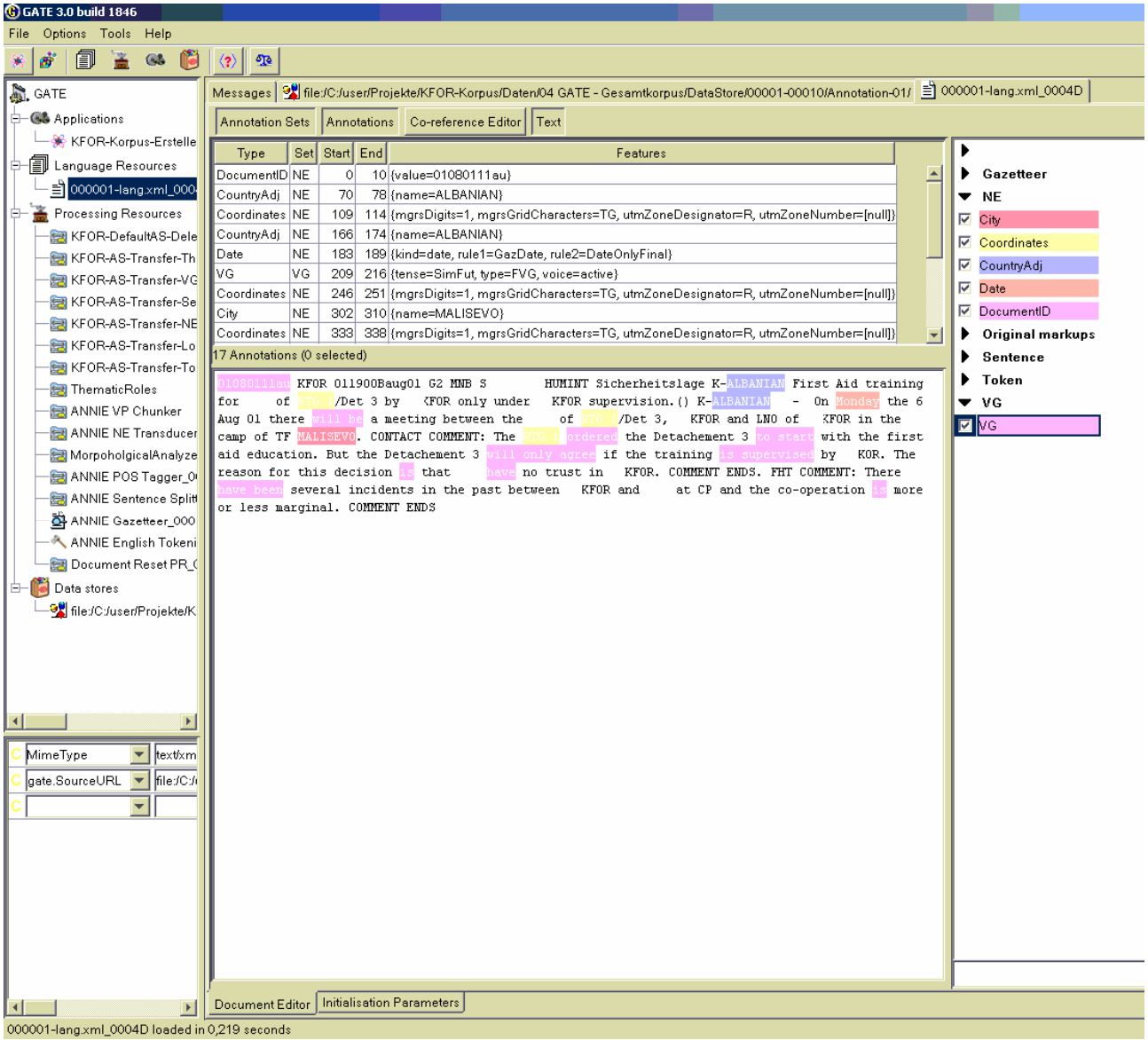
- KFOR corpus: Approx. 4.000 Military HUMINT reports from the KFOR deployment of the Bundeswehr are the basis for the realization of the KFOR corpus.
- specialized micro text corpus
- 886,000 tokens
- The corpus is classified.

- Annotation layers:
  - Original markups: parts of the message which are already formatted (e.g. addressee, topic, source)
  - Token: annotations from the tokenizer and the POS tagger
  - Gazetteer: expressions identified through lists of names
  - Sentence: sentences begin and end, markers of comments
  - Named entities
  - Verb phrases

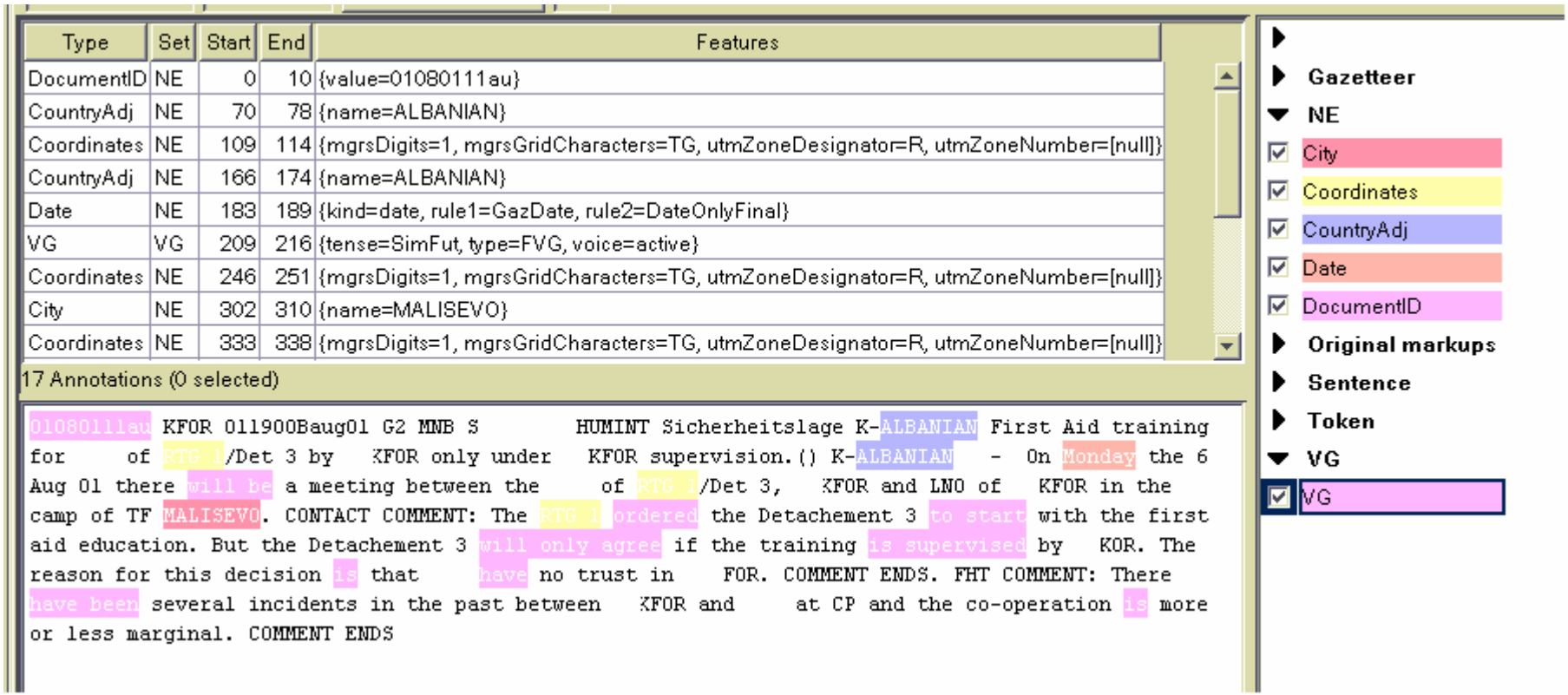
- During the creation of the corpus a first version of the annotations was produced automatically.
- These annotations were then manually checked and corrected.
- For both working-steps GATE was used.
- The corpus is represented in
  - the GATE-specific format,
  - the GATE-specific format in XML,
  - an stand-off annotation format.

- For the following purposes the military messages and the corpus can be used:
  - They represent the basis for the construction of the IE component. The lexicon and the transducers were optimized towards the corpus.
  - The performance of the IE of the ZENON prototype can evaluated quantitatively relative to the corpus.
  - The corpus can be used for other research objectives (e.g., complexity of nominal phrases, word sense disambiguation, machine learning of grammatical structures, etc.).

Dr. M. Hecking


GATE 3.0 build 1846
File Options Tools Help
GATE
Applications
KFOR-Korpus-Erstelle
Language Resources
000001-lang.xml_000
Processing Resources
KFOR-DefaultAS-Dele
KFOR-AS-Transfer-Th
KFOR-AS-Transfer-VG
KFOR-AS-Transfer-Se
KFOR-AS-Transfer-NE
KFOR-AS-Transfer-Lo
KFOR-AS-Transfer-To
ThematicRoles
ANNIE VP Chunker
ANNIE NE Transducer
MorpoholgicalAnalyze
ANNIE POS Tagger_0
ANNIE Sentence Split
ANNIE Gazetteer_000
ANNIE English Tokeni
Document Reset PR_(
Data stores
file:/C:/user/Projekte/K
MimeType
text/xm
gate.SourceURL
file:/C:/
Messages
file:/C:/user/Projekte/KFOR-Korpus/Daten/04 GATE - Gesamtkorpus/DataStore/00001-00010/Annotation-01/
000001-lang.xml_0004D
Annotation Sets
Annotations
Co-reference Editor
Text
Type Set Start End Features
DocumentID NE 0 10 {value=01080111au}
CountryAdj NE 70 78 {name=ALBANIAN}
Coordinates NE 109 114 {mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}
CountryAdj NE 166 174 {name=ALBANIAN}
Date NE 183 189 {kind=date, rule1=GazDate, rule2=DateOnlyFinal}
VG VG 209 216 {tense=SimFut, type=FVG, voice=active}
Coordinates NE 246 251 {mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}
City NE 302 310 {name=MALISEVO}
Coordinates NE 333 338 {mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}
17 Annotations (0 selected)
01080111au KFOR 011900Baug01 G2 MNB S HUMINT Sicherheitslage K-ALBANIAN First Aid training for of /Det 3 by (FOR only under KFOR supervision.() K-ALBANIAN - On Monday the 6 Aug 01 there will be a meeting between the of /Det 3, KFOR and LNO of KFOR in the camp of TF MALISEVO. CONTACT COMMENT: The ordered the Detachement 3 to start with the first aid education. But the Detachement 3 will only agree if the training is supervised by KOR. The reason for this decision is that have no trust in KFOR. COMMENT ENDS. FHT COMMENT: There have been several incidents in the past between KFOR and at CP and the co-operation is more or less marginal. COMMENT ENDS
Gazetteer
NE
City
Coordinates
CountryAdj
Date
DocumentID
Original markups
Sentence
Token
VG
VG
Document Editor
Initialisation Parameters
000001-lang.xml_0004D loaded in 0,219 seconds

Dr. M. Hecking



Type	Set	Start	End	Features
DocumentID	NE	0	10	{value=01080111au}
CountryAdj	NE	70	78	{name=ALBANIAN}
Coordinates	NE	109	114	{mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}
CountryAdj	NE	166	174	{name=ALBANIAN}
Date	NE	183	189	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
VG	VG	209	216	{tense=SimFut, type=FVG, voice=active}
Coordinates	NE	246	251	{mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}
City	NE	302	310	{name=MALISEVO}
Coordinates	NE	333	338	{mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}

17 Annotations (0 selected)

01080111au KFOR 011900Baug01 G2 MNB S HUMINT Sicherheitslage K-ALBANIAN First Aid training for of RTG 1/Det 3 by KFOR only under KFOR supervision.() K-ALBANIAN - On Monday the 6 Aug 01 there will be a meeting between the of RTG 1/Det 3, KFOR and LNO of KFOR in the camp of TF MALISEVO. CONTACT COMMENT: The RTG 1 ordered the Detachement 3 to start with the first aid education. But the Detachement 3 will only agree if the training is supervised by KOR. The reason for this decision is that have no trust in FOR. COMMENT ENDS. FHT COMMENT: There have been several incidents in the past between KFOR and at CP and the co-operation is more or less marginal. COMMENT ENDS

- Gazetteer
- NE
  - City
  - Coordinates
  - CountryAdj
  - Date
  - DocumentID
- Original markups
- Sentence
- Token
- VG
  - VG